Bayesian Approaches for Information Borrowing, Sample Size Estimation and Adaptive Decisions: BART and BESS

Yuan Ji, PhD

October 3, 2025

University of Chicago
Dept. of Public Health Sciences
Collaboration with Dehua Bi (Stanford), Jiaxin Liu (Bayesoft)Yunxuan Zhang (UChicago), Gu Mi & Ji Lin at Sanofi

Motivation

Motivation

- External data such as real-world data (RWD) or prior trial data can reduce the need for large control arms in RCTs.
- When borrowing information from external data, three main issues need to be considered:
  - the similarity between the external data and the trial population in terms of the covariates;
  - how to adjust for unmeasured confounders;
  - the effective sample size of the external data
- Methods matching or weighting the propensity scores are commonly used to adjust for measured confounders.
- The Meta-Analytic Predictive (MAP) prior is a common approach for incorporating historical controls partially address the degree of borrowing caused by additional unmeasured confounding.

MAP Recap

- **Setup:** For $j = 1, \ldots, J$, observe study means $\bar{y}_j$ with known SEs s_j:

$$\bar{y}_j \mid \theta_j \sim N(\theta_j,\, s_j^2),$$

$$\theta_j \mid \mu, \tau^2 \sim N(\mu,\, \tau^2),$$

$$(\mu, \tau^2) \sim p(\mu, \tau^2).$$

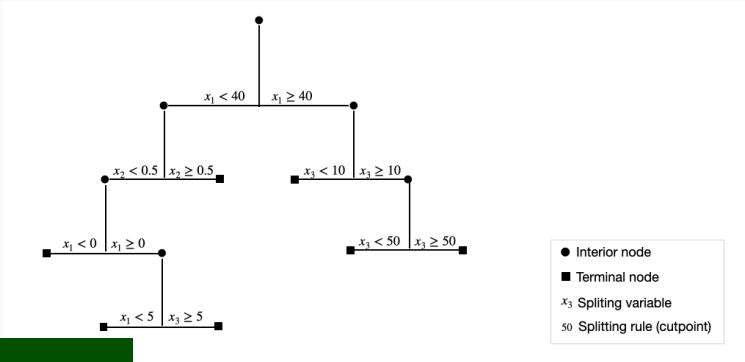
- Here, θ_j = study-specific mean; μ = grand mean; τ^2 = between-study variance.

BART-powered Borrowing from External Data

BART = Bayesian Additive Regression Trees

Put simply: "A collection of regression trees, each updated via Bayesian rules, whose predictions are then summed into one final estimate."

Here's what a typical **regression tree** looks like:



BART Recap

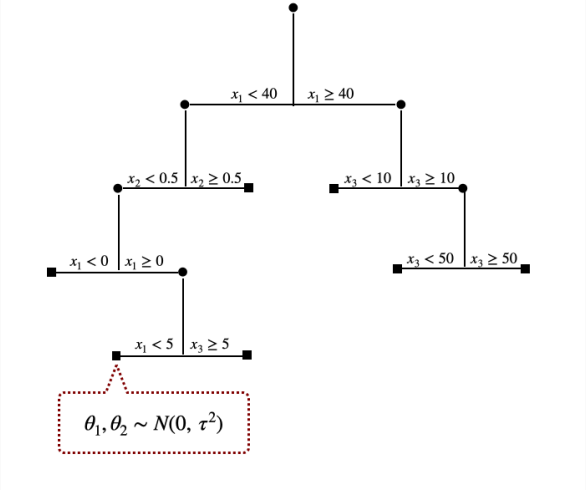
Given $(y_i, \boldsymbol{x}_i)$, BART models

$$y_i = \sum_{h=1}^{H} g(\boldsymbol{x}_i; T_h, \theta_h) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- T_h: binary regression tree defined by
  - Internal nodes: each splits on a covariate x at cutpoint c
  - Terminal (leaf) nodes: final regions l where the node mean θ_{hl} applies, routing each observation i
- Terminal node means: $\theta_{hl} \overset{iid}{\sim} \mathcal{N}\big(0, \hat{\tau}^2\big)$, where $\hat{\tau}^2$ is *determined from the data*.

MAP-BART: Illustration

Idea: BART with MAP Embed Terminal Nodes



MAP-BART: Model specification

Let $D_i \in \{1, 2\}$ denote the data source indicator for observation i:

$$D_i = \begin{cases} 1, & \text{RCT control,} \\ 2, & \text{RWD control.} \end{cases}$$

Likelihoods:

- Outcome model (Gaussian)

$$Y_{1i} \mid X_{1i}, \{T_h, \Theta_h\}_{h=1}^{H}, \sigma_1^2 \sim \mathcal{N}\Big(\sum_{h=1}^{H}\sum_{l=1}^{L_h} (\mathbf{1}_{X_{1i} \in \text{ Node}_{hl}}) \cdot \theta_{1hl}, \sigma_1^2\Big)$$

$$Y_{2i} \mid X_{2i}, \{T_h, \Theta_h\}_{h=1}^{H}, \sigma_2^2 \sim \mathcal{N}\Big(\sum_{h=1}^{H}\sum_{l=1}^{L_h} (\mathbf{1}_{X_{2i} \in \text{ Node}_{hl}}) \cdot \theta_{2hl}, \sigma_2^2\Big)$$

MAP-BART: Model specification cont.

- Outcome model (Log-normal with right censoring)

$$Y_i \text{ : event time,} \quad C_i \text{ : censoring time}$$

$$f(\cdot) \text{ and } S(\cdot) \text{ are the Normal pdf/survival function.}$$

$$\log(Y_{1i}) \mid X_{1i}, \{T_h, \Theta_h\}_{h=1}^{H}, \sigma_1^2 \sim \begin{cases} f\big(\log Y_{1i} \,\big|\, \mu_{1i}, \sigma_1^2\big), & Y_{1i} \le C_{1i} \text{ (observed)} \\ S\big(\log C_{1i} \,\big|\, \mu_{1i}, \sigma_1^2\big), & Y_{1i} > C_{1i} \text{ (censored)} \end{cases}$$

$$\log(Y_{2i}) \mid X_{2i}, \{T_h, \Theta_h\}_{h=1}^{H}, \sigma_2^2 \sim \begin{cases} f\big(\log Y_{2i} \,\big|\, \mu_{1i}, \sigma_2^2\big), & Y_{2i} \le C_{2i} \text{ (observed)} \\ S\big(\log C_{2i} \,\big|\, \mu_{2i}, \sigma_2^2\big), & Y_{2i} > C_{2i} \text{ (censored)} \end{cases}$$

$$\mu_{gi} = \sum_{h=1}^{H} \sum_{l=1}^{L_h} \mathbf{1}_{\{X_{gi} \in \mathsf{Node}_{hl}\}} \cdot \theta_{ghl}, \quad g = 1, 2.$$

MAP-BART: Model specification cont.

Priors:

1. Prior for trees: $\{T_h\}_{h=1}^{H}$
   - a node at depth d is split with probability $p_{\text{split}}(d) = \alpha(1+d)^{-\beta}$;
   - the splitting variable is chosen uniformly from the p predictors;
   - the cut-point is then selected uniformly from the admissible cut-points of that variable.
2. Prior for terminal node means: $\{\Theta_h\}_{h=1}^{H}$

$$\theta_{1hl}, \theta_{2hl} \mid T_h, \tau_{hl}^2 \sim \mathcal{N}\Big(\mathbf{0},\ \tau_{hl}^2\Big),$$

$$\tau_{hl}^2 \sim \text{Inv-}\chi^2(\nu, s^2).$$

Node–specific variance τ_{hl}^2 allows borrowing tailored to each terminal node.

Average Treatment Effect (ATE)

- **Definition**

$$\mathsf{ATE} = \mathbb{E}_{\mathbb{X}}\Big[\mathbb{E}_{\mathbb{Y}}\big[Y(1) \mid X = x\big]\Big] - \mathbb{E}_{\mathbb{X}}\Big[\mathbb{E}_{\mathbb{Y}}\big[Y(0) \mid X = x\big]\Big]$$

- **Gaussian**: $Y(a) \mid X = x \sim \mathcal{N}(\mu_a(x), \sigma^2)$

$$\mathsf{ATE} = \mathbb{E}_{\mathbb{X}}\big[\mu_1(x) - \mu_0(x)\big].$$

- **Log-normal**: $\log Y(a) \mid X = x \sim \mathcal{N}(\mu_a(x), \sigma^2)$

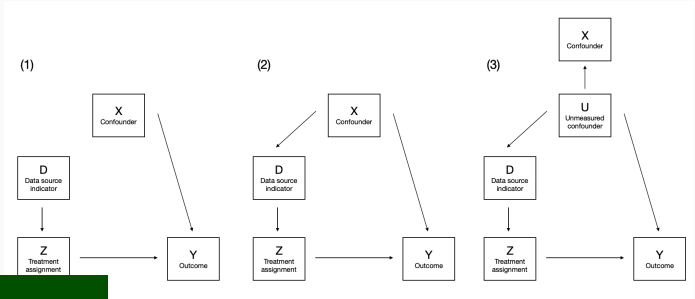
$$\mathsf{ATE} = \frac{q_{0.5}\left(\mathbb{E}_{\mathbb{X}}[S_1(t \mid x)]\right)}{q_{0.5}\left(\mathbb{E}_{\mathbb{X}}[S_0(t \mid x)]\right)} := \frac{\inf\left\{t : \mathbb{E}_{\mathbb{X}}[S_1(t \mid x)] \leq 0.5\right\}}{\inf\left\{t : \mathbb{E}_{\mathbb{X}}[S_0(t \mid x)] \leq 0.5\right\}}$$

[1] $S_a(t \mid x) = \Pr\big\{\log Y(a) > t \mid X = x\big\}$ denotes the survival function.
[2] $q_{0.5}(F)$ denotes the 0.5 quantile (median) of distribution F.

Simulation Study

Simulation Design

- **Endpoint:** Continuous (Gaussian) and time-to-event (log-normal)
- **Scenarios: Varying confounding mechanisms**
  1. No confounding, $D \sim \text{Bernoulli}(0.5)$.
  2. Measured confounding: $\Pr(D = 1)$ depends on X.
  3. Unmeasured confounding: $\Pr(D = 1)$ depends on U, while $\text{cor}(X, U) = \rho \in \{0, 0.25, 0.5, 0.75\}$.



Simulation Design cont.

- **Replicates:** $R = 500$
- **Sample sizes per replicate**
  - RCT treatment vs. control: 200 vs. 100
  - RWD control: 300
  - Prior ESS constrained to < 200 (see next slide)
- **Outcome model (Gaussian)**

$$Y_i \mid D_i = 1 \text{ (RCT)} : Y_i = \alpha_{\text{RCT}} + \beta_{\text{RCT}}^T X_i{}^{1} + \delta\, Z_i + \varepsilon_i,$$

$$Y_i \mid D_i = 0 \text{ (RWD)} : Y_i = \alpha_{\text{RWD}} + \beta_{\text{RWD}}^T X_i{}^{1} + \varepsilon_i,$$

with $\alpha_{\text{rct}} = 0.15,\ \alpha_{\text{rwd}} = 0.5,\ \beta_{\text{rct}} = (-1, -0.2, \ldots, -0.2),\ \beta_{\text{rwd}} = \beta_{\text{rct}},\ \delta = 2,\ \varepsilon_i \sim N(0, \sigma^2),\ \sigma = 0.4$ for rct, $\sigma = 1$ for rwd.

[1] For Scenarios 1 and 2, $X_i = X_i$; for Scenario 3, $X_i = U_i$.

Simulation Design cont.

- **Outcome model (log normal)**

$$Y_i \mid D_i = 1 \text{ (RCT)} : \log Y_i = \alpha_{\text{RCT}} + \beta_{\text{RCT}}^T X_i + \delta\, Z_i + \varepsilon_i,$$

$$Y_i \mid D_i = 0 \text{ (RWD)} : \log Y_i = \alpha_{\text{RWD}} + \beta_{\text{RWD}}^T X_i + \varepsilon_i,$$

with $\alpha_{\text{rct}} = 0.15,\ \alpha_{\text{rwd}} = 0.5,\ \beta_{\text{rct}} = (-1, -0.2, \ldots, -0.2),\ \beta_{\text{rwd}} = \beta_{\text{rct}},\ \delta = \log(1.25),\ \varepsilon_i \sim N(0, \sigma^2),\ \sigma = 0.4$ for rct, $\sigma = 1$ for rwd.

Censoring mechanism:

- *Dropout censoring:* $C_i \sim \text{Exponential}(\lambda_c)$
- *Staggered entry:* $E_i \sim \text{Uniform}(0,\ 1.25)$ (max enrollment at 1.25 yrs)
- *Admin censoring:* $C_{\text{admin},i} = 2.5 - E_i$ (study ends at 2.5 yrs)
- *Observed time and event indicator:*

$$Y_i^{\text{obs}} = \min\{Y_i,\ C_i,\ C_{\text{admin},i}\}, \qquad \Delta_i = \mathbf{1}\left\{Y_i \le \min(C_i,\ C_{\text{admin},i})\right\}$$

Evaluation Metrics

- **Bias**, **Posterior SD** and **Root Mean Squared Error (RMSE)** in estimating the Average Treatment Effect (ATE) *per replicate*

  Let $\delta = \mathbb{E}[Y(1) - Y(0)]$ be the true ATE, and $\delta^{(m)}$ the estimate in the iteration $m = 1, \ldots, M$. We compute:

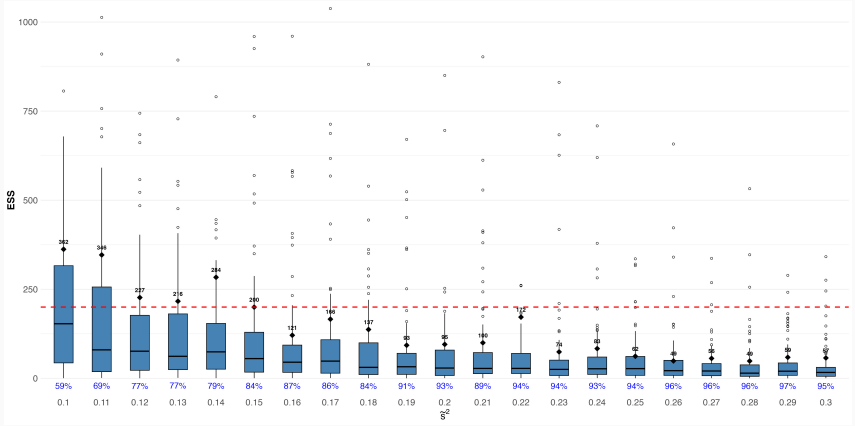
$$\textbf{Bias} = \underbrace{\frac{1}{M}\sum_{m=1}^{M} \delta^{(m)}}_{\text{Posterior mean } \bar{\delta}} - \delta,$$

$$\textbf{SD} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M} \left(\delta^{(m)} - \bar{\delta}\right)^2},$$

$$\textbf{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M} \left(\delta^{(m)} - \delta\right)^2}.$$

Prior Effective Sample Size (ESS) Tuning

- The hyperparameter s^2 was set to 0.25 so that 95% of simulated prior ESS values < 200.



Results

Results by Outcome Type

Scenario 1: No confounding — Gaussian

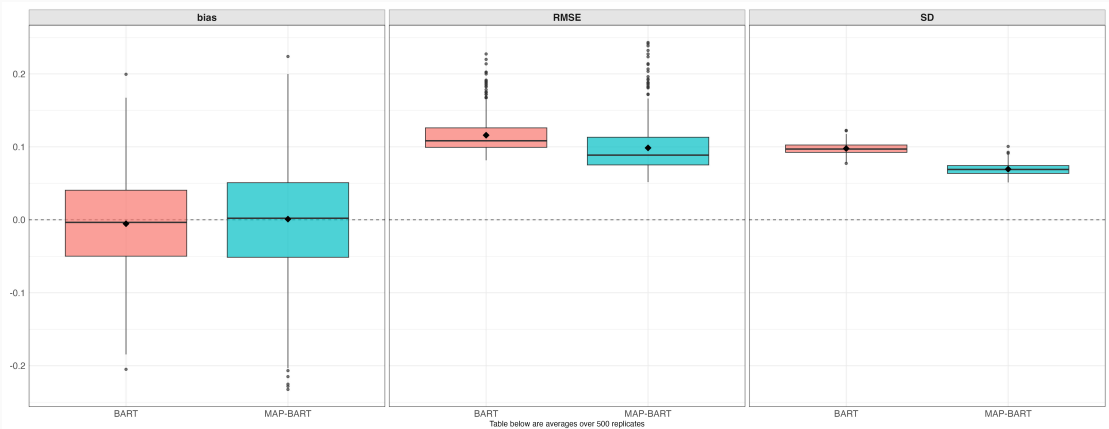


Table below are averages over 500 replicates

Method	Bias	SD	RMSE	CI_length	CI_coverage
BART	-0.01	0.10	0.12	0.38	1.00
MAP-BART	0.00	0.07	0.10	0.27	0.93

Results by Outcome Type

Scenario 1: No confounding — Log-normal

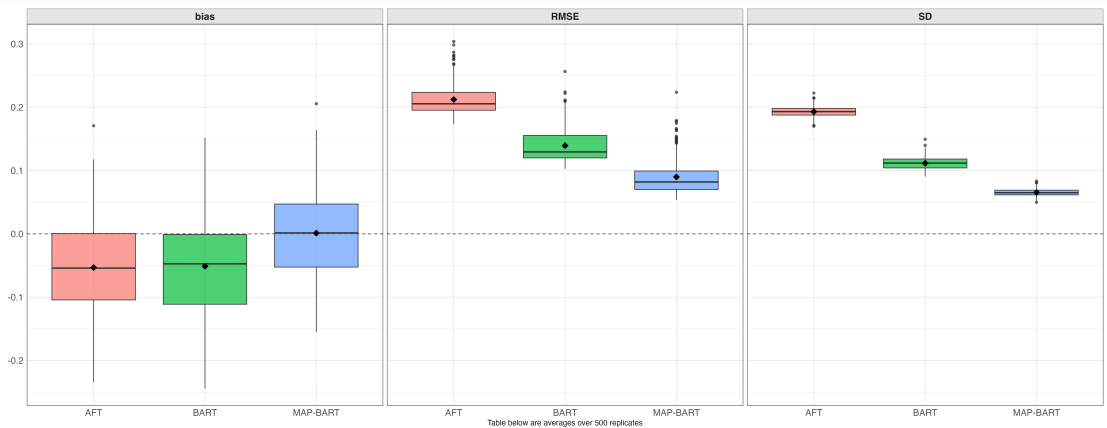


Table below are averages over 500 replicates

Method	Bias	SD	RMSE	CI_length	CI_coverage
AFT	-0.05	0.19	0.21	0.76	1.00
BART	-0.05	0.11	0.14	0.44	0.99
MAP-BART	0.00	0.07	0.09	0.26	0.96

Results by Outcome Type

Scenario 2: Measured confounding — Gaussian

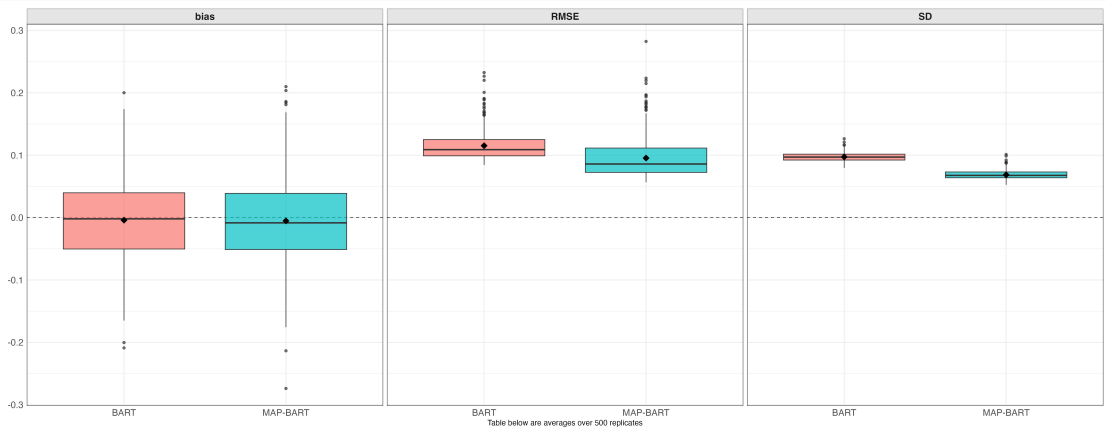


Table below are averages over 500 replicates

Method	Bias	SD	RMSE	CI_length	CI_coverage
BART	0.00	0.10	0.12	0.38	1.00
MAP-BART	-0.01	0.07	0.10	0.27	0.93

Results by Outcome Type

Scenario 2: Measured confounding — Log-normal

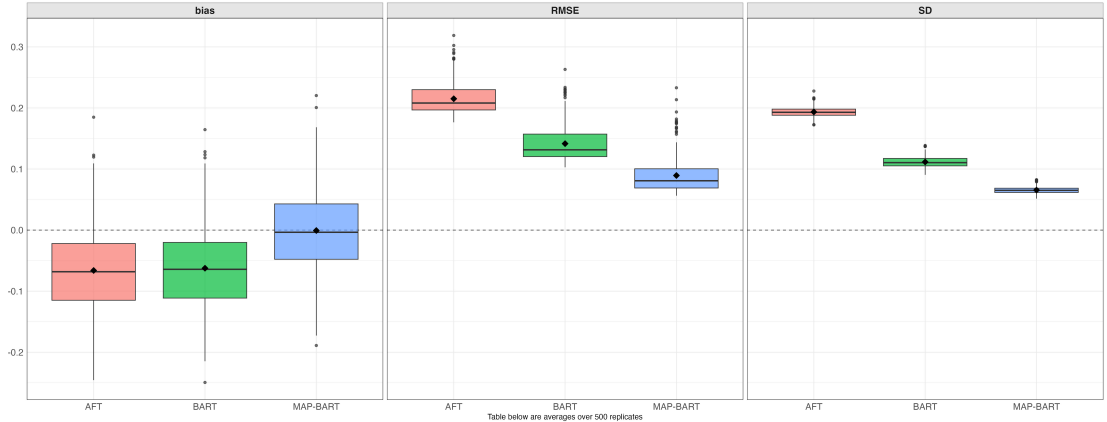
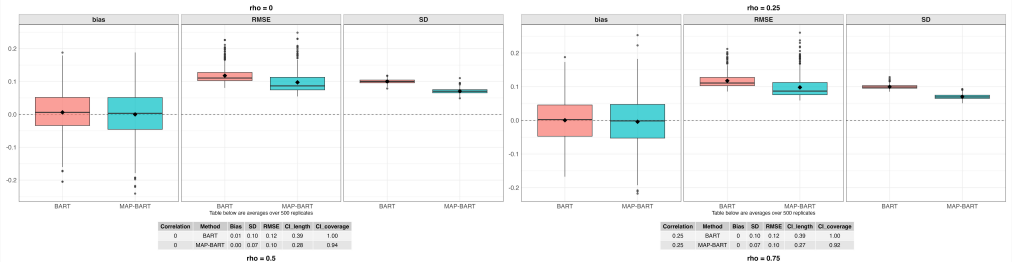


Table below are averages over 500 replicates

Method	Bias	SD	RMSE	CI_length	CI_coverage
AFT	-0.07	0.19	0.22	0.76	1.00
BART	-0.06	0.11	0.14	0.44	0.98
MAP-BART	0.00	0.07	0.09	0.26	0.94

Results by Outcome Type

Scenario 3: Unmeasured confounding — Gaussian



Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0	BART	0.01	0.10	0.12	0.39	1.00
0	MAP-BART	0.00	0.07	0.10	0.28	0.94

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.25	BART	0	0.10	0.12	0.39	1.00
0.25	MAP-BART	0	0.07	0.10	0.27	0.92

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.5	BART	0	0.10	0.12	0.39	0.99
0.5	MAP-BART	0	0.07	0.10	0.27	0.95

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.75	BART	0	0.10	0.12	0.38	0.99
0.75	MAP-BART	0	0.07	0.10	0.27	0.91

Results by Outcome Type

Scenario 3: Unmeasured confounding — Log-normal

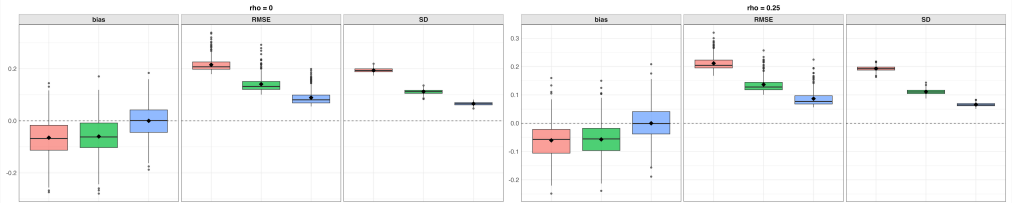


Table below are averages over 500 replicates

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0	AFT	-0.06	0.19	0.22	0.76	1.00
0	BART	-0.06	0.11	0.14	0.44	0.98
0	MAP-BART	0.00	0.07	0.09	0.26	0.95

Table below are averages over 500 replicates

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.25	AFT	-0.06	0.19	0.21	0.76	1.00
0.25	BART	-0.06	0.11	0.14	0.43	0.99
0.25	MAP-BART	0.00	0.07	0.09	0.26	0.95

Table below are averages over 500 replicates

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.5	AFT	-0.07	0.19	0.22	0.76	1.00
0.5	BART	-0.06	0.11	0.14	0.44	0.97
0.5	MAP-BART	0.00	0.07	0.09	0.26	0.93

Table below are averages over 500 replicates

Correlation	Method	Bias	SD	RMSE	CI_length	CI_coverage
0.75	AFT	-0.06	0.19	0.21	0.76	1.00
0.75	BART	-0.06	0.11	0.14	0.44	0.99
0.75	MAP-BART	0.00	0.07	0.09	0.26	0.94

Conclusion

Conclusion

- MAP-BART marries flexible non-parametric regression with principled historical borrowing.
- Node-specific variances and t-approximation enable efficient tree moves.
- Ongoing: real oncology dataset application.

References

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. The Annals of Applied Statistics, 4(1), 266–298. https://doi.org/10.1214/09-AOAS285

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Guo, P., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics, 70(4), 1023–1032. https://doi.org/10.1111/biom.12209

Zhou, T., & Ji, Y. (2021). Incorporating external data into the analysis of clinical trials via Bayesian additive regression trees. Statistics in Medicine, 40(30), 6782–6795. https://doi.org/10.1002/sim.9196

BESS: Bayesian Estimator of Sample Size

Sample Size Estimation Is an Educated Guess

There is no data Y in sample size estimation.

Frequentist Estimation framework uses the only probability measure P_Y.

- Given true θ, the uncertainty is about data Y described by the probability measure P_Y.
- Sample size $n \to Y(n) \to P_{Y(n)}$.
- Under $P_{Y(n)}$, require $P_{Y(n)}(T(Y) > t_0)$ is small under H_0 (α) and large under H_1 $(1-\beta)$.

Frequentist: Guess the truth, under which to trade off sample size and type I/II error rates (based on the probability measure of data $P_{Y(n)}$);

Standard Sample Size Estimation (SSE)

Consider comparing two proportions θ_0 and θ_1, the standard SSE

Difference in Proportions: The minimum difference to be detected between the two proportions θ^* - a smaller difference requires a larger sample size.

Significance Level (α): the Type I error rate.

Power (1-β): $1-\beta$; β the Type II error rate.

True Proportions: The assumed proportions θ_0 and θ_1.

Then assume 1:1 allocation, sample size per arm is

$$n = \frac{(z_\alpha + z_\beta)^2}{(\theta_1 - \theta_0 - \theta^*)^2}[\theta_1(1-\theta_1) + \theta_0(1-\theta_0)]. \quad (1)$$

SSE Statement – hard to interpret for non-statisticians

Sample size formula:

$$n = \frac{(z_\alpha + z_\beta)^2}{(\theta_1 - \theta_0 - \theta^*)^2}[\theta_1(1-\theta_1) + \theta_0(1-\theta_0)].$$

Statement 1: At a significance level of α, with a clinically minimum effect size θ^*, n subjects are needed to achieve $(1-\beta)$ power when the response rates for the treatment and control arms are θ_1 and θ_0.

α & β: Hard to interpret to investigators; build for multiple experiments but not a single trial

True θ_1 and θ_0: Can never verify if the assumed values are true or not

Promise If investigators are willing to spend n subjects-worth resource, statisticians promise α and $(1-\beta)$ as return. – back to this later...

Three Pillars of BESS

Consider comparing two resp. rates θ_0 and θ_1.

Hypotheses: let $\theta = (\theta_1 - \theta_0)$. Consider

$$H_0 : \theta \leq \theta^* \text{ vs. } H_1 : \theta > \theta^*, \tag{2}$$

Sample size n The number of subjects in data $\boldsymbol{y}_n = \{y_{ij}\}$, $i = 1, ..., n$.

Evidence $e(\boldsymbol{y}_n)$ Assumed evidence from data, e.g., $e = \bar{y}_1 - \bar{y}_0$

Confidence Posterior probability $\Pr(H_1|\boldsymbol{y}_n)$.

{Decision rule (Müller et al., 2014):} Reject H_0 and accept H_1 if $\Pr(H_1|\boldsymbol{y}_n) > c$.

BESS Needs An Hierarchical Model

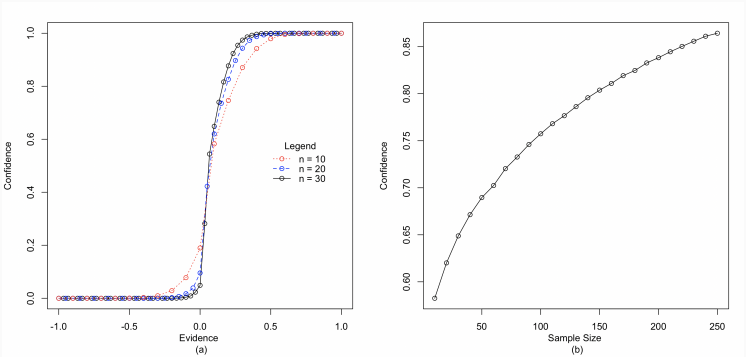
Consider binary, continuous, and count data. Bayesian hierarchical models

Outcome type	Parameter θ_j	Likelihood $f(.)$	Prior distribution $\pi_j(\tilde{\boldsymbol{\theta}})$
Binary	Response rate	Bern(θ_j)	Beta(a, b)
Continuous	Mean response	$N(\theta_j, \sigma^2)$, σ known	$N(a, b)$
Count-data	Event rate	Poi(θ_j)	Gamma(a, b)

Table 1: Summary of parameter, likelihood function, and prior distribution for different data types.

Case 1: For desired evidence $e(\boldsymbol{y}_n)$ and confidence cutoff c, find n satisfies $\Pr(H_1|\boldsymbol{y}_n) > c$.

Testing (Superiority) $H_0 : \theta \leq 0.05$ vs. $H_1 : \theta > 0.05$, $\theta = (\theta_1 - \theta_0)$.



Case 2: For fixed sample size n, evidence $e(\boldsymbol{y}_n)$ from data leads to confidence higher than c, where $\Pr(H_1|\boldsymbol{y}_n) > c$.

Testing (non-inferiority) $H_0 : \theta_L - \theta_H \leq \theta^*$ vs. $H_1 : \theta_L - \theta_H > \theta^*$,

	Sample size $n = 20$									
Evidence $\bar{y}_L - \bar{y}_H$	≤ -0.20	-0.15	-0.10	-0.05	0.00	0.05	0.10	0.15	0.20	≥ 0.25
Confidence c	< 0.05	0.12	0.28	0.50	0.62	0.74	0.84	0.90	0.94	> 0.95

Table 2: List of various evidence and confidence for $\theta^* = -0.05$ with $n = 20$ patients per arm.

Coherent – A Necessary Condition

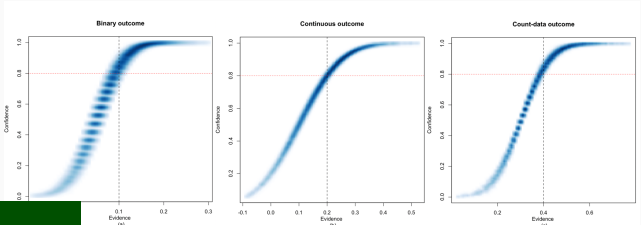
The Bayesian design (BESS) and Bayesian inference is coherent

Design: for assumed evidence e, a sample size of n will provide confidence c that the alternative hypothesis is true.

Data: Denote $\boldsymbol{y}_n^\star$ the observed data, $e^\star$ the observed evidence, and $c^\star = \Pr(H_1|\boldsymbol{y}_n^\star)$ the posterior probability of H_1 conditional on the observed data $\boldsymbol{y}_n^\star$.

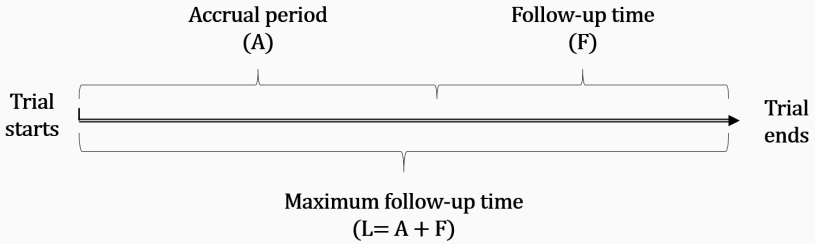
Weak Coherence: if $e^\star > e$, $c^\star > c$.

Strong Coherence: Confidence c increases with e (or n) for fixed n (or e).



Frequentist Methods for Survival

Two -Arm Superiority Design



- A: Accrual period; F: Follow-up period; $L = A + F$: max follow-up time
- n_j: sample size in arm j, allocation ratio $k = \frac{n_1}{n_0}$
- λ_0: hazard rate for control; λ_1: hazard rate for treatment

Standard Sample Size Estimation (Schoenfeld (1981))

Hypotheses:

$$H_0 : \lambda_1/\lambda_0 \geq r \quad vs. \quad H_1 : \lambda_1/\lambda_0 < r$$

- Significance level α: Type I Error Rate
- Power $(1-\beta)$: β: Type II Error Rate

Number of Events

$$n_d = \frac{[(1+k)(z_{1-\alpha} + z_{1-\beta})]^2}{k(\log(\frac{\lambda_1}{\lambda_0}) - \log(r))^2}$$

Sample Size

$$n = \frac{n_d}{p}, n_0 = \frac{n}{1+k}, n_1 = \frac{kn}{1+k},$$

where combined probability of event $p = \frac{p_0}{1+k} + \frac{kp_1}{1+k}$

Standard Sample Size Estimation (Schoenfeld (1981, 1983))

Following Schoenfeld (1983), p_0 and p_1 can be obtained in two ways under the exponential survival assumption.

Method 1 (Simpson approximation, conservative). For the control arm (λ_0):

$$p_0 \;=\; 1 \;-\; \frac{1}{6}\Big\{ e^{-\lambda_0 F} + 4\, e^{-\lambda_0 (F+0.5A)} + e^{-\lambda_0 (F+A)} \Big\}.$$

The event probability in the treatment arm is then

$$p_1 \;=\; 1 - (1 - p_0)^{\lambda_1/\lambda_0}$$

Method 2 (closed-form, less conservative). A direct integration under the exponential model and uniform accrual yields

$$p_j \;=\; 1 - \frac{1}{A\lambda_j}\Big(e^{-\lambda_j F} - e^{-\lambda_j (F+A)} \Big) \qquad j = 0, 1.$$

Standard Sample Size Estimation (Lachin-Foulkes (1986))

According to the central limit theorem,

$$\sqrt{n_j}(\hat{\lambda}_j - \lambda_j) \to_d N(0, \sigma^2(\lambda_j)),$$

where $\hat{\lambda_j} = \frac{\sum_{i=1}^{n_j} \delta_{ij}}{\sum_{i=1}^{n_j} y_{ij}}$ and $\sigma^2(\lambda_j) = \lambda_j^2 \left[1 + \frac{e^{-\lambda_j L}}{A\lambda_j}(1 - e^{\lambda_j A})\right]^{-1}$.

Let $\varepsilon = \lambda_0 - \lambda_1$, the problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0 : \varepsilon \leq \delta \quad \text{vs.} \quad H_1 : \varepsilon > \delta,$$

where δ is the superiority or noninferiority margin. The test statistic is :

$$T = (\hat{\lambda}_0 - \hat{\lambda}_1 - \delta) \left[\frac{\sigma^2(\hat{\lambda}_0)}{n_0} + \frac{\sigma^2(\hat{\lambda}_1)}{n_1}\right]^{-1/2}$$

Under the assumption that $n_1 = kn_0$, we have

$$n_0 = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\varepsilon - \delta)^2} \left[\frac{\sigma^2(\lambda_1)}{k} + \sigma^2(\lambda_0)\right].$$

SSE Statement – hard to interpret for non-statisticians

Number of events formula:

$$n_d = \frac{[(1+k)(z_{1-\alpha} + z_{1-\beta})]^2}{k(\log(\frac{\lambda_1}{\lambda_0}) - \log(r))^2}.$$

Statement 1: At a significance level of α, with a clinically minimum effect threshold r, n_d events are needed to achieve $(1-\beta)$ power when the hazard rates for the treatment and control arms are λ_1 and λ_0.

α & β: Hard to interpret to investigators; build for multiple experiments but not a single trial

True λ_1 and λ_0: Can never verify if the assumed values are true or not

Methodology

Notations and Time Definitions

- t_{ij}: enrollment time of patient i in arm j
- T_{ij}: event time; y_{ij}: follow-up time
- δ_{ij}: event indicator, if $\delta_{ij} = 1$, the event is observed, $y_{ij} = T_{ij}$; else if $\delta_{ij} = 0$, the follow-up time is right censored, and $y_{ij} = L - t_{ij}$.

Under **exponential model:** $T_{ij} \sim \text{Exp}(\lambda_j)$, the **likelihood contribution for subject** i **in arm** j**:**

$$(\lambda_j e^{-\lambda_j y_{ij}})^{\delta_{ij}} (e^{-\lambda_j y_{ij}})^{1-\delta_{ij}}$$

We use $\boldsymbol{y} = \{\delta_{ij}, y_{ij}; i = 1, \cdots, n_j, j = 0, 1\}$ to denote the data. Use $d_j = \sum_{i=1}^{n_j} \delta_{ij}$ and $V_j = \sum_{i=1}^{n_j} y_{ij}$ to denote number of events and follow up time under arm j, respectively.

Three Pillars of BESS-Surv

Consider comparing two hazard rates λ_0 and λ_1.

Hypotheses:

$$H_0 : \lambda_1/\lambda_0 \geq r \quad vs. \quad H_1 : \lambda_1/\lambda_0 < r$$

Sample Sizes d_0, d_1 The numbers of events in both arms

Evidence $e(\boldsymbol{y})$ Summary of treatment effects, e.g., $e = (\hat{\lambda}_0(\boldsymbol{y}), \hat{\lambda}_1(\boldsymbol{y}))$; $\hat{\lambda}_0(\boldsymbol{y}) = \frac{d_0}{V_0}$, $\hat{\lambda}_1(\boldsymbol{y}) = \frac{d_1}{V_1}$ are estimated hazard rates using trial data

Confidence Posterior probability $\Pr(H_1|\boldsymbol{y})$,

Decision rule (Müller et al., 2014): Reject H_0 and accept H_1 if $\Pr(H_1|\boldsymbol{y}) > c$.

Posterior Probability of H_1

$$T_{ij} \sim f(\lambda_j) = \mathsf{Exp}(\lambda_j)$$

$$\lambda_0, \lambda_1 \mid \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, H_l \sim \pi(\lambda_0, \lambda_1 \mid \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, H_l) I(\lambda_0, \lambda_1 \in H_l) \qquad l = 0, 1$$

$$\Pr(H = H_1) = q,$$

Priors: $\lambda_j \sim \mathsf{Gamma}(\alpha_j, \beta_j)$, joint prior under hypothesis H_l, define $C_l = \frac{1}{\int\int_{\lambda_0, \lambda_1 \in H_l} \pi(\lambda_0; \alpha_0, \beta_0)\pi(\theta_1; \alpha_1, \beta_1) d\lambda_0 d\lambda_1}$: Posterior probability:

$$\begin{aligned}\Pr(H = H_1 \mid \boldsymbol{y}) =& \frac{p(\boldsymbol{y} \mid H = H_1)\Pr(H = H_1)}{p(\boldsymbol{y} \mid H = H_1)\Pr(H = H_1) + p(\boldsymbol{y} \mid H = H_0)\Pr(H = H_0)} \\ =& \frac{C_1 \Pi_1 q}{C_1 \Pi_1 q + C_0 \Pi_0 (1 - q)}\end{aligned}$$

with:

$$\Pi_l = \int\int \pi(\lambda_0; \alpha_0 + d_0, \beta_0 + V_0)\pi(\lambda_1; \alpha_1 + d_1, \beta_1 + V_1) I(\lambda_0, \lambda_1 \in H_l) d\lambda_0 d\lambda_1$$

Sample Size Estimation Algorithm

1. Assume the observed hazard rates in the two arms $e = (\hat{\lambda}_0(\boldsymbol{y}), \hat{\lambda}_1(\boldsymbol{y}))$, in which $\hat{\lambda}_j(\boldsymbol{y}) = \frac{d_j}{V_j}$.
2. Input the confidence c, prior probability q and d_{max}.
3. Set $d_0 = 1$. We impose the restriction that $d_1, d_0 \geq 1$.
4. For each value of $d_0, d_1 \in \{1, \ldots, d_{\text{max}}\}$,
   4.1 Compute $V_0 = d_0/\hat{\lambda}_0(\boldsymbol{y})$ and $V_1 = d_1/\hat{\lambda}_1(\boldsymbol{y})$,
   4.2 Calculate $\Pr(H = H_1 \mid \boldsymbol{y})$.
   4.3 Calculate $n_j = \left\lceil \frac{d_j}{\pi_j} \right\rceil$.
5. Compare
   5.1 If $\Pr(H = H_1 \mid \boldsymbol{y}) \geq c$ and $n_1/n_0 = k$, record n_0 and n_1
6. Among all the recorded (n_0, n_1), the final sample size is smallest n_0 and n_1 from Step 5.

where

$$1 - \frac{1}{\quad}\left(e^{-\lambda_j F} - e^{-\lambda_j (F+A)}\right) \qquad j = 0, 1.$$

Coherent – Property and Condition

The Bayesian design (BESS) and Bayesian inference is coherent

Design: For assumed evidence e, number of events of d_0, d_1 will provide confidence c that the alternative hypothesis is true.

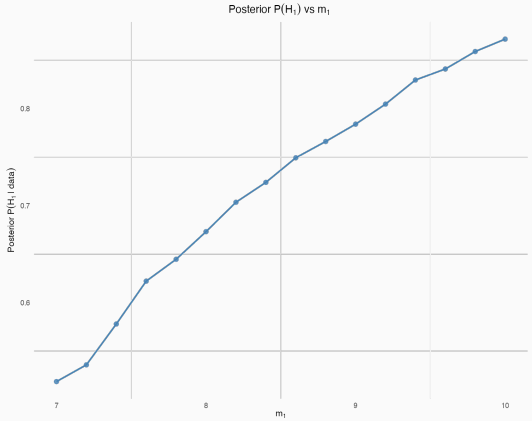
Data: Denote $\boldsymbol{y}^\star$ the observed data, $e^\star$ the observed evidence, and $c^\star = \Pr(H_1|\boldsymbol{y}^\star)$ the posterior probability of H_1 conditional on the observed data $\boldsymbol{y}^\star$.

Weak Coherence: if $e^\star > e$, $c^\star > c$.

Strong Coherence: Confidence c increases with e (or d_0, d_1) for fixed d_0, d_1 (or e).

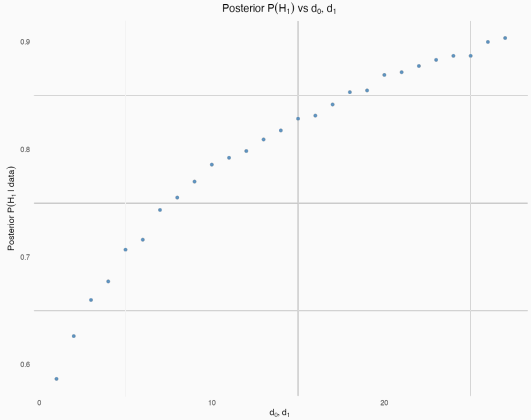
Coherent – For Fixed Number of Events

As for the survival condition, the evidence: $e = (\hat{\lambda}_0(\boldsymbol{y}), \hat{\lambda}_1(\boldsymbol{y}))$, we define an increased evidence as fix $\hat{\lambda}_0(\boldsymbol{y})$, and increase the median survival for treatment $m_1 = \frac{\log(2)}{\hat{\lambda}_1(\boldsymbol{y})}$.



Coherent – For Fixed Evidence

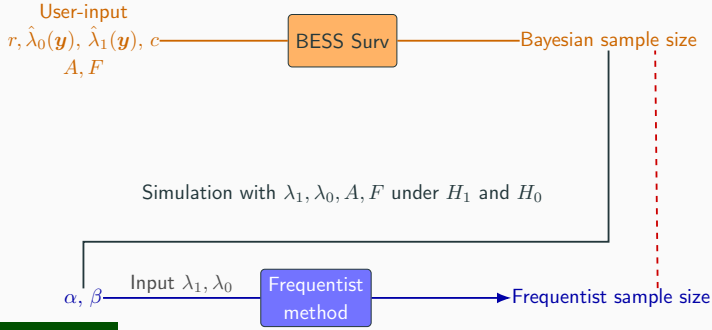
We fix evidence $e = (\hat{\lambda}_0(\boldsymbol{y}), \hat{\lambda}_1(\boldsymbol{y}))$, and increase the number of events.



Simulation Results

Simulation Results

In the simulation setup, we fix $\hat{\lambda}_0(\boldsymbol{y}) = \log(2)/7$, and vary $\hat{\lambda}_1(\boldsymbol{y})$ along with different values of c. Under the alternative hypothesis H_1, we set the true rates as $\lambda_0 = \log(2)/7$ and $\lambda_1 = \log(2)/9$; under the null hypothesis H_0, both rates are equal: $\lambda_0 = \lambda_1 = \log(2)/7$. We also set the threshold for the hazard ratio comparison at $r = 1$.



Simulation Results

Table 3: Simulation results comparing BESS and standard methods.

$\hat{\lambda}_1(\boldsymbol{y})$	c	α	$1-\beta$	BESS	Schoenfeld1	Schoenfeld2	Lachin
log(2)/8	0.7	0.297	0.680	96	134	93	94
	0.8	0.200	0.765	236	325	225	228
	0.9	0.096	0.874	558	795	549	557
log(2)/9	0.7	0.290	0.525	34	51	35	36
	0.8	0.195	0.516	72	108	75	76
	0.9	0.098	0.504	158	226	156	159
log(2)/10	0.7	0.303	0.459	18	23	16	16
	0.8	0.196	0.424	40	59	41	41
	0.9	0.097	0.373	86	127	88	89

Trial Example

An Example

Hypotheses:

$$H_0 : \lambda_1/\lambda_0 \geq r \quad vs. \quad H_1 : \lambda_1/\lambda_0 < r$$

- Median Survival for two arm: $m_0 = 7$, $m_1 = 10$
- Corresponding $(\hat{\lambda}_0(\boldsymbol{y}), \hat{\lambda}_1(\boldsymbol{y})) : (0.099, 0.069)$
- Accrual $A = 12$, Follow-up $F = 8$, Equal allocation
- Set confidence = 0.9, $r = 1$, the estimated sample size using BESS is: $n_0 = n_1 = 41$

Simulation Setup

Two scenarios: the Null Scenario and Alternative Scenario. For each of the scenario,we generate the hazard rate using the following rationales:

Step 1: Sample λ_0 from LogNormal distribution: $\lambda_0 \sim \text{LogNormal}(a, b)$.

Step 2: Sample λ_1 based on λ_0:

$$\log(\lambda_1) \mid \log(\lambda_0) \sim \begin{cases} \text{TN}(a, b, \log(\lambda_0 \times r), \text{Inf}) & H_0 \\ \text{TN}(a_1, b_1, \text{-Inf}, \log(\lambda_0 \times r)) & H_1, \end{cases}$$

where a, b, a_1, b_1 are calculated parameters under assuming the 2.5%tile and 97.5%tile for m_1 is 8 and 13, and 4 and 9 for m_0, respectively. We generate 1,000 trials under each scenario by setting $n_0 = n_1 = 41$.

Given the generated data for each trial, calculate $\Pr(H = H_1 \mid \boldsymbol{y})$. If it is greater than or equal to $c = 0.9$, reject H_0 and accept H_1.

Simulation Results

Based on the 2,000 simulated trials (1,000 each for the Null or Alternative Scenario), we report the performance of BESS-Surv computed sample size based on four different metrics. The simulation results are summarized in Table 4.

Table 4: Simulation results. FDR: False Discovery Rate; FOR: False Omission Rate; FPR: False Positive Rate; FNR: False Negative Rate.

	FDR	FOR	FPR	FNR
Error rate	0.057	0.240	0.042	0.303

Overall, the sample size calculated using BESS leads to reasonable error rates.

Discussion

Benefits & Future work

- Sample size re-estimation can be incorporated if needed.
- No adjustment to the Type I error rate is required, offering flexibility for multi-arm designs.
- May integrate historical data into the prior for sample size estimation.

Reference

[1] Bi, D., & Ji, Y. (2024). A Bayesian Estimator of Sample Size. https://arxiv.org/abs/2404.07923

[2] Schoenfeld, D. A. [1983]. Sample-size formula for the proportional-hazards regression model, Biometrics pp. 499–503.

[3] Chow, S.-C., Shao, J., Wang, H. and Lokhnygina, Y. [2017]. Sample size calculations in clinical research, Chapman and Hall/CRC.